IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 10, 2019, accepted July 29, 2019, date of publication July 31, 2019, date of current version August 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2932146

Multimodal Camera-Based Gender Recognition Using Human-Body Image With Two-Step Reconstruction Network

NA RAE BAEK, SE WOON CHO, JA HYUNG KOO, NOI QUANG TRUONG, AND KANG RYOUNG PARK

Division of Electronics and Electrical Engineering, Dongguk University, Seoul 04620, South Korea

Corresponding author: Kang Ryoung Park (parkgr@dongguk.edu)

This work was supported in part by the National Research Foundation of Korea (NRF) of the Ministry of Education through the Basic Science Research Program under Grant NRF-2018R1D1A1B07041921, in part by the NRF of the Ministry of Science and ICT through the Basic Science Research Program under Grant NRF-2019R1F1A1041123, and in part by the NRF of the Korean Government, MSIT, through the Bio and Medical Technology Development Program under Grant NRF-2016M3A9E1915855.

ABSTRACT With the recent development of intelligent surveillance systems, the importance of research study on gender recognition of people at a distance is also on the rise. The existing gender recognition technologies studies have used high-resolution facial images captured from the front at a short distance, which showed high performance. However, intelligent surveillance systems in actual real environments have difficulty in detecting the faces of people because they use images captured from a distance. Moreover, in the case of back-view images, gender recognition based on the facial image is impossible because the face cannot be detected. Thus, gender recognition using the full-body human-body images of people is being studied but its performance is low owing to problems such as low resolution, motion blur, and optical blur. Furthermore, the performance of gender recognition using only visible-light cameras is limited owing to illumination variations, shadow, and the type of clothes and accessories. To solve these problems, remote body-shape-based recognition was performed by sequentially using two convolutional neural networks which improved the resolution of visible-light images. In addition, the degradation of recognition performance owing to various factors (e.g., illumination, shadow, and the type of clothes and accessories) was prevented by combining a visible-light camera with an infrared camera, and the scalability was enhanced using various heterogeneous cameras. The higher performance of the proposed method compared with that of other methods was verified through a comparative experiment using the open database of Sun Yat-sen University multiple modality Re-ID (SYSU-MM01) and the Dongguk body-based gender database (DBGender-DB2) that has been built by us.

INDEX TERMS Gender recognition, visible-light and infrared cameras, image reconstruction, human-body image, convolutional neural network.

I. INTRODUCTION

Many studies have been conducted on gender recognition based on human-body images, and these studies have been used in many applications such as improving the retrieval accuracy of gender-based human searching, demographic data collection, security surveillance, customer statistics collection, criminal matching, and list monitoring. Furthermore, biometric recognition or face recognition can improve the accuracy and speed of recognition using gender-information-based pre-classification.

Most of the existing gender recognition studies use high-resolution facial images [1]–[3]. However, images captured in an intelligent surveillance environment have poor quality as shown in Figure 1 because of the long distance between people and the camera, and the face image cannot even be detected from the back view of people. Moreover, people's cooperation is required to use their facial images for gender recognition. Using human body images can be an alternative solution because people's images acquired from the surveillance environment contain human body information such as

The associate editor coordinating the review of this manuscript and approving it for publication was Marco Anisetti.

FIGURE 1. Example images captured in intelligent surveillance systems. The upper images are human body images and the lower images are the face parts cropped from the human body images.

body movement and body appearance. However, study on gender recognition based on human body image has not progressed considerably yet because of a few challenges. First, gender recognition based on human body image uses body shape, hair style, and type of clothes and accessories as cues. However, peoples of same gender may wear different styles of clothes. Furthermore, unisex clothes can appear similar even between different genders. The same is true for hair styles. Second, gender recognition must be robust to human poses, variations, and occlusions [4]. Third, human images are captured in various environments: illumination variations, various backgrounds, motion blur, optical blur, etc. [5]. Owing to these challenges, there are considerably fewer gender recognition studies based on human-body image than those based on face or gait.

Existing studies on gender recognition based on human-body image only use visible-light cameras. However, when gender recognition is performed only using images acquired from visible-light cameras, the recognition accuracy is low owing to some factors such as background, clothes, and accessories. To solve this problem, we use infrared (IR) cameras based on the temperature difference between human body and background regions to decrease the influence of such factors. We improve the gender recognition performance by combining a visible-light camera, which describes details on human body, and an IR camera, which describes human body shape.

Furthermore, most images in a surveillance environment have low resolution, because those images are captured from a distance. Low-resolution visible-light images degrade the gender recognition performance because there is less information on people's shape. In this study, we improve the gender recognition performance through a deep residual network (ResNet) after applying single-image super-resolution (SISR) to low-resolution visible-light images. In addition, we show the higher performance of the proposed method in comparison with that of the existing gender recognition based on human-body image through a comparison experiment using the open databases, Sun Yat-Sen University Multiple Modality Re-ID (SYSU-MM01) [7] and Dongguk Body-based Gender Database (DBGender-DB2) [8].

In Section II, the advantages and disadvantages of various existing gender recognition methods are compared and analyzed.

II. RELATED WORKS

Various existing studies on gender recognition are introduced in Ng *et al.* [9]. Existing gender recognition studies can be largely divided into face-based, body-movement-based, and body-appearance-based methods. The face-based methods [3], [10]–[14] use information that can distinguish between men and women in clear facial images captured at a short distance. Furthermore, studies to obtain this information for distinguishing gender, i.e., better features, have been conducted. There are various feature extraction methods not only for handcrafted features such as image texture features [10], principal component analysis (PCA) [11], Haar-like feature [12], and local binary pattern (LBP) [3], but also for deep-learned features such as deep convolutional neural network [13] and hyperface [14].

However, as shown in Figure 1, it is difficult to use low-resolution facial images captured at a distance as in intelligent surveillance systems. Moreover, occluded images or non-frontal images cannot be used because the entire faces are not detected.

Regarding these problems of intelligent surveillance systems, body-movement-based method or body-appearance-based method using the body information are being studied. The body-movement-based method involves performing gender recognition using the gait information of people in successive images. Lee and Grimson [15] divided a human silhouette into seven ellipses and extracted moment-based features from each ellipse. However, this method has the disadvantage of high feature dimension because it uses the entire silhouette. Hence, Yu *et al.* [16] decreased the feature dimension by using noise-robust gait energy images (GEI) [17] and average silhouettes. A study observed that, although the GEI is powerful in representing human gait because it is robust to pre-processing noise, it is sensitive to changes in viewpoint or pose [19]. To address this problem, Lu *et al.* [18] proposed a method using cluster-based averaged gait image (C-AGI) after clustering gait sequences into similar views or poses. C-AGI features are represented in many views and poses, resulting in inter-class variations generate. They achieved a high performance by introducing sparse reconstruction-based metric learning (SRML), which minimizes intra-class distance and maximizes inter-class distance. These body-movement-based methods [15], [16], [18] can use low-resolution images captured at a distance and also use back view images that do not have human face information. However, they have a disadvantage that they take a long time to obtain gait images with human body movement information such as silhouette, GEI, and C-AGI from successive images. Also, when objects do not move, it is difficult to obtain the body movement information. Moreover, segmentation information is required to separate walking objects from the

background, and it cannot be used for occluded images with cut-off legs or faces of people.

Considering these problems of body-movement-based methods, body-appearance-based methods with relatively short processing time using single images have been studied. Body-appearance-based methods can be divided into single-camera-based methods and multimodal-camera-based methods. Most of the existing studies are single-camera-based methods using images acquired through visible-light cameras, and they can be divided into handcrafted-feature-based methods and deep-feature-based methods. The studies in [4], [5], [20], [21] used the histogram of oriented gradient (HOG) among the handcrafted features. Cao *et al.* [4] studied gender recognition through body appearance, i.e., the full-body images of people. After dividing a full-body image into patch units, they extracted features from each patch through the HOG [22] and performed classification through adaptive boosting (AdaBoost) [23]. However, this study did not use color information properly. It is described in [7] that color information is important in gender recognition. Therefore, Collins *et al.* [5] used features that combine hue-saturation-value (HSV) color histogram features with pixelHOG (PiHOG), which is a dense HOG computed from a custom edge map. Bourdev *et al.* [20] extracted features based on random patches called poselets [25] consisting of people's HOG features, color histogram, and skin-mask features. Although the method proposed in [20] is more robust to pose and occlusion than previous studies, it has a limitation in that a heavily annotated training dataset is required. Guo *et al.* [21] proposed biologically inspired features (BIFs) as a new handcrafted-feature-based method. BIFs are extracted through a Gabor filter and the feature dimension was reduced through manifold learning such as PCA and locality-sensitive discriminant analysis (LSDA). The gender was recognized after view classification (e.g., frontal view, back view, mixed view), rather than simple classification into men and women. BIF with LSDA showed a high recognition performance in frontal or back view, and BIF with PCA showed a high recognition performance in mixed views.

This handcrafted-feature-based method requires a separate classifier. However, the deep-feature-based method does not require a separate classifier because it uses a convolutional neural network (CNN), which is a single framework where feature extraction and classification are integrated, and the features are also automatically trained from training data. Krizhevsky *et al*. [26] presented impressive recognition results in object recognition through CNN. Based on this, CNN has been used for many pattern recognition problems and has shown excellent classification performance. Ng *et al.* [27] proposed a gender recognition method using a CNN consisting of two convolution layers, two subsampling layers, and one fully-connected layer. It was experimentally verified that the method using CNN achieved a performance higher than or similar to that of the handcrafted-feature-based method. Antipov *et al.* [28] proposed a gender recognition method based on compact CNN and Mini-CNN. They showed through Mini-CNN and fine-tuned AlexNet [26] that deep features have a higher recognition performance in heterogeneous data than HOG features. Previous studies used global patches, i.e., total images of people, but Ng *et al.* [29] trained local patches in each CNN and combined them with the results of global-patch-based CNN. They classified local patches into top, middle, and bottom patches of human bodies and experimented with single patches. They observed that the top patches showed the highest performance, followed by the middle patches. Hence, they achieved a high performance by combining the two high-performance local patches, top and middle patches. They used visual geometry group (VGG) Net-19 [30] for the architecture. Cai *et al.* [31] proposed HOG-assisted deep feature learning (HDFL), which combines deep features with weighted HOG features, which are handcrafted features. After extracting features simultaneously from input images, they combined the two features in the fusion layer to obtain more discriminative features. They performed gender recognition through a Softmax classifier from these features. Raza *et al.* [32] proposed a stacked sparse autoencoder (SSAE) in which the result of a sparse autoencoder (SAE) is input to the next SAE. To train the SSAE, a parsing phase is applied in which people and background are binarized and the background is removed. Subsequently, gender recognition is performed by applying this image to the SSAE with Softmax classifier.

This single-camera-based method has a limited recognition performance owing to various factors such as illumination variations, background, shade, and type of clothes and accessories. Thus, Nguyen *et al.* [33] combined a visible-light camera with an IR camera based on temperature differences between human body and background area, which is less affected by these factors and achieved a higher performance than previous studies that only used a visible-light camera. Furthermore, they used two feature extraction methods, HOG and multi-level LBP (MLBP), and HOG showed a higher performance. Nguyen *et al.* [33] simply used visible-light and thermal images, and the background region affected the HOG features, which lowered the performance. Hence, Nguyen *et al.* [34] proposed a weighted HOG method based on the characteristic of thermal image that the human body region is brighter than the background and on the image quality assessment that assesses the quality of sub-blocks. They created a standard deviation map (STD map) by measuring quality using the mean and standard deviation of the gray levels of sub-blocks of the thermal image based on the characteristic that the background region appears darker in the thermal image. The STD map indicates the probability of belonging to the background or human body regions, and the weight of the background region was lowered, and the weight of the human body region was increased. Consequently, the background region was less affected during the HOG feature extraction and the performance improved. However, these gender recognition methods [33], [34] have a limited recognition performance because they use a predesigned handcrafted

TABLE 1. Comparisons of previous and proposed methods for gender recognition.

Category				Advantage	Disadvantage
Face-based [3, 10-14]				High recognition performance because high-resolution images are used. Not affected by the clothing styles of people	Subjects' cooperation is required to use the facial images of people. It requires relatively high-quality images captured at a short distance and cannot be used for images with occluded faces. Back views cannot be used for recognition.
Body-based	Body-movement-based [15, 16, 18]			Back views can also be used for recognition. Relatively low-quality images captured at a distance can be used as well.	It takes much time to process because it is based on multiple images. Segmentation information and walking subject from the background are required. It is difficult to use for occluded images with cut-off legs and faces of people.
	Body-appearance-based	Single-camera-based [4, 5, 20, 21, 27-29, 31, 32]		Front, side, or back view images captured at a distance can be used, and segmentation information for the foreground is not necessary.	Recognition accuracy is decreased by such factors as illumination variations, shade, background, clothes, and accessories because only single-camera images are used.
		Multimodal-camera-based	Without image reconstruction [33-35]	Front, side, or back view images captured at a distance can be used, and segmentation information for the foreground is not necessary. Less affected by such factors as illumination variations, shade, background, clothes, and accessories because it combines thermal camera and visible-light camera information.	Recognition accuracy is low because low-resolution noisy and blurred images captured at a distance are used.
			Two-step image reconstruction (proposed method)	Front, side, or back view images captured at a distance can be used, and segmentation information for the foreground is not necessary. Less affected by such factors as background, clothes, and accessories because it is combined with information from various IR cameras. The recognition accuracy is improved through image denoising and super-resolution reconstruction.	It takes time to process two-step image reconstruction.

feature extractor. Thus, Nguyen *et al.* [35] performed gender recognition based on the deep features of CNN. They extracted and fused the features of visible-light image and thermal image based on AlexNet. They removed noise and reduced the feature dimensions using PCA. Subsequently, they performed classification using a support vector machine. Although gender recognition based on human body was not performed, previous research proposed a new tracking system which aims at fusing the information from RGB and infrared modalities based on machine learning model for object tracking [57]. In [58], authors proposed an infrared tracking system where information from RGB-modality was exploited to assist the infrared object tracking. In [59], authors proposed a feature representation and fusion model in order to fuse the feature representation of the object in RGB and infrared modalities with the dual-camera systems for capturing RGB and infrared videos for object tracking.

Most of these studies on gender recognition based on human body image have a limited recognition performance owing to motion blur, optical blur, and sensor noise as they use low-resolution images captured at a distance. Furthermore, most studies only consider visible-light images, which limit performance owing to illumination variations, background, clothes, and accessories. Considering these problems of existing studies, we propose a gender recognition method based on human body images captured at a distance by fusing visible-light images to which a two-step image reconstruction based on deep CNN is applied and various IR images. Table 1 compares the advantages and disadvantages of the proposed method and existing methods on gender recognition.

III. CONTRIBUTIONS

Compared with previous studies, this study has the following four contributions:

- This is the first study that improved the performance of body-image-based gender recognition through deep-CNN-based denoising and super-resolution reconstruction (SR).
- This study performed two-step image reconstruction to reduce blur and noise when restoring low-resolution visible-light images through SR. Through various experiments, we proved that two-step image reconstruction shows a better performance than performing SR alone.
- To reduce the training time, we only trained the CNN used in gender recognition without that for 2-step image construction. In addition, through filter images and feature maps, the characteristic differences between visible-light and NIR images for gender recognition were analyzed.
- The scalability of the body-image-based gender recognition was improved through various combinations of

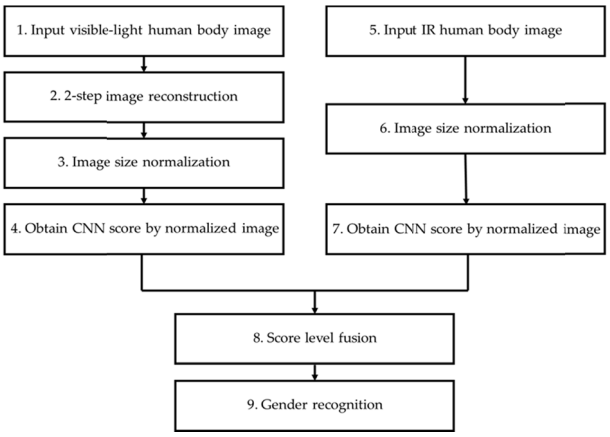


FIGURE 2. Flowchart of the proposed method.

heterogeneous cameras (combination of visible-light camera and near-infrared (NIR) camera, and combination of visible-light camera and thermal camera). Furthermore, for fair assessment of performances of other studies, the gender classification label information of our trained CNN models and SYSU-MM01 database were revealed through [24].

IV. PROPOSED METHOD

A. OVERVIEW OF PROPOSED METHOD

Figure 2 shows the overall procedure of the proposed method. First, human body images are acquired through visible-light and IR cameras in an intelligent surveillance environment (Steps (1) and (5) in Figure 2). The acquired human body images are low-resolution images with severe blur and noise because the subjects are people walking at a distance. In the preprocessing step, the resolution of the visible-light image is enhanced by applying two-step image reconstruction (Step (2) in Figure 2). The CNN architecture used in the two-step image reconstruction is detailed in Section IV-B. After size normalization of this reconstructed image (Step (3) in Figure 2), it is input to the deep residual network and a score of gender recognition is obtained (Step (4) in Figure 2). Furthermore, the input IR (NIR or thermal) image undergoes size normalization (Step (6) in Figure 2) and is input to the deep residual network and a score of gender recognition is obtained (Step (7) in Figure 2). The scores obtained from the visible-light and IR (NIR or thermal) images are fused (Step (8) in Figure 2), and gender recognition for the input image is performed (Step (9) in Figure 2). The deep residual network for obtaining the score of gender recognition is described in Section IV-C, and the gender recognition based on the score level fusion is described in Section IV-D.

In our researches, we used the two open databases of SYSU-MM01 and DBGender-DB2 for experiments. In the first database, visible-light and NIR light images were included whereas visible-light and thermal images were included in the second database. We can regard both NIR and thermal images as infrared (IR) images, and there exist

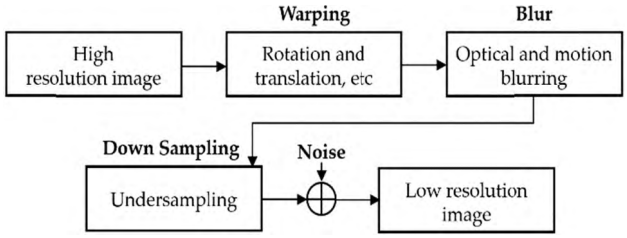


FIGURE 3. Process of transformation from a high-resolution image to a low-resolution image.

FIGURE 4. Example images obtained by applying CNN-based SR reconstruction through VDSR to original images. The left images are the original images and the right images are the resulting images of SR reconstruction.

the differences between NIR and thermal images. Thermal image is captured based on thermal radiation that is shown in the wavelength of 8–12 μm [65]. Therefore, it is called long wavelength IR (LWIR) light. NIR image is captured based on light whose wavelength is much shorter (0.75–1.4 μm) than LWIR.

Thermal image can be acquired without additional illuminator whereas the acquisition of NIR image usually requires additional NIR illuminators, especially in night. Therefore, the capturing distance of NIR camera is limited due to the limitation of the illumination distance. In addition, the impact of absorption and scattering of fog is known to be less severe in the LWIR light than NIR one [65].

B. TWO-STEP CNN-BASED IMAGE RECONSTRUCTION

Images captured at a distance have negative effects on gender recognition owing to motion blur, optical blur, noise, and low resolution. Consequently, the recognition performance is decreased because pixel information related to shape, hair style, clothes, and accessories, which are important for gender recognition using visible-light images, is lost. To solve this problem, this study performed two-step CNN-based image reconstruction for low-resolution human body images captured at a distance. In the first step, image denoising is performed using image restoration CNN (IRCNN). In the second step, image SR reconstruction is performed using very deep convolutional networks SR (VDSR). Section IV-B-1 describes the general image SR reconstruction process, and the two CNNs are detailed in Sections IV-B-2 and IV-B-3.

1) DESCRIPTION OF IMAGE SR RECONSTRUCTION

SISR is a method of obtaining a high-resolution image from a low-resolution image. In general, a high-resolution image is transformed into a low-resolution image through warping,

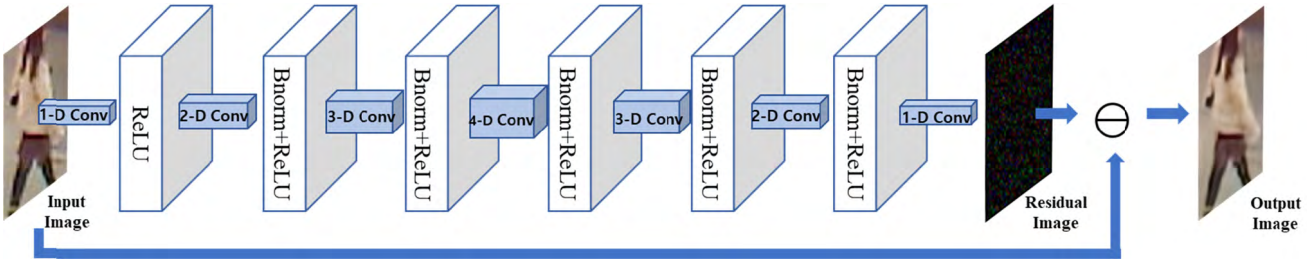


FIGURE 5. Architecture of IRCNN. Bnorm and ReLU indicate batch normalization and rectified linear unit, respectively. N-D Conv indicates N dilated convolutional layer.

blurring, down sampling, and noise addition as shown in Figure 3. This SISR for a low-resolution image is considered an ill-posed problem [49], [50]. This process can be expressed as Equations (1) and (2) [49], [50]:

$$y_k = D(B_k(W_k(x)))+n_k \quad (1)$$

$$y_k = Tx = H_k x + n_k = DB_k W_k x + n_k \quad (2)$$

where y_k is the low-resolution image, D is the sub-sampling matrix, B_k is the blur matrix, W_k is the warping matrix, x is the original high-resolution image, and n_k is the noise vector. Furthermore, D,B_k, and W_k can be represented by H_k, and T becomes the reconstruction matrix. The existing image reconstruction methods [38]–[44] determined the optimal T matrix based on Equation (2). In addition, image reconstruction was performed through interpolation, neighbor embedding, and sparse coding, but the performance improvement was limited. Consequently, Dong et al. [45] proposed a CNN-based image SR method. As this has a higher reconstruction performance than existing methods, there are many studies on this CNN-based image SR reconstruction. Based on the results of these studies, we improved the image reconstruction performance by using two CNNs. These two CNNs are detailed in Sections IV-B-2 and IV-B-3.

2) THE 1ST STEP OF IMAGE RECONSTRUCTION

In this study, instead of directly applying the CNN-based SR reconstruction such as VDSR to the input image, IRCNN [36] was first applied before the CNN-based SR reconstruction. This is because, when the CNN-based SR reconstruction is simply applied to low-resolution images, boundary artifacts are generated as shown in Figure 4. These boundary artifacts lower the recognition performance because the shape information of people also plays an important role in gender recognition. Therefore, we used IRCNN first, and VDSR was applied as the second step. As shown in [36], IRCNN outperforms the traditional denoising methods such as block-matching and 3D filtering (BM3D) [60], weighted nuclear norm minimization (WNNM) [61], Thompson-Nicola regional district (TNRD) [62], multi-layered perceptron (MLP) [63], and color version of block-matching and 3D filtering (CBM3D) [64] in terms of denoising accuracy. Based on these results, IRCNN is used in our research. Figure 5 shows the architecture of IRCNN. IRCNN expanded the receptive filter while maintaining the 3×3 convolutional filter by using a dilated convolutional filter.

Furthermore, it used small-sized patches for training so that the CNN could recover more boundary information. The study in [36] applied additive Gaussian noise with noise level σ after first applying the blur kernel to the image when training IRCNN. It improved performance by applying the noise levels step by step.

Furthermore, as shown in Figure 5, IRCNN is a residual-learning-based method and works by subtracting the residual image with unnecessary information from the low-resolution image by learning the residual image in the low-resolution image. When applying IRCNN in this study, we do not train separately with the experimental data of this study, but we use pre-trained model. This is because there are no pairs of noisy and denoised images among the human body images captured in the surveillance environment used in this study, which makes training difficult.

3) THE 2ND STEP OF IMAGE RECONSTRUCTION

The image denoised by IRCNN described in Section IV-B-2 is input to the VDSR. Figure 6 and Table 3 show the architecture of the VDSR, which is also a residual-learning-based method. It trains a residual image that contains the high-resolution information to be added to the low-resolution image and adds the residual image into the input low-resolution image to obtain the final high-resolution image [37]. As it trains such that the residual image, instead of the high-resolution image, would become the output of CNN, training is performed well even with a deep structure of 20 layers and at relatively high learning rates. As shown in the residual image in Figure 6, the VDSR further strengthens the shape information of people. Consequently, it denoises the image captured at a distance in the first step of image reconstruction and adds the shape information of the human body image in the second step of image reconstruction, thus improving the overall image resolution. As with IRCNN, the pre-trained model for VDSR in [37] was used without separate training with the experimental data in this study. This is because there are no pairs of low- and high-resolution images among the human body images captured in the surveillance environment used in this study, which makes training difficult.

TABLE 2. Detail descriptions of IRCNN structure.

Layer type	Number of filters	Size of feature map (width × height × channel)	Size of kernel (width × height)	Number of strides	Number of paddings
Input layer [image]		W × H × 3			
1-D Conv 1 (1st convolutional layer)	64	W × H × 64	3×3	1×1	1×1
ReLU 1		W × H × 64			
2-D Conv 2 (2nd convolutional layer)	64	W × H × 64	3×3	1×1	2×2
Bnorm + ReLU 2		W × H × 64			
3-D Conv 2 (3rd convolutional layer)	64	W × H × 64	3×3	1×1	3×3
Bnorm + ReLU 3		W × H × 64			
4-D Conv 4 (4th convolutional layer)	64	W × H × 64	3×3	1×1	4 x 4
Bnorm + ReLU 4		W × H × 64			
3-D Conv 5 (5th convolutional layer)	64	W × H × 64	3×3	1×1	3×3
Bnorm + ReLU 5		W × H × 64			
2-D Conv 6 (6th convolutional layer)	64	W × H × 64	3×3	1×1	2×2
Bnorm + ReLU 6		W × H × 64			
1-D Conv 7 (7th convolutional layer)	64	W × H × 3	3×3	1×1	1×1

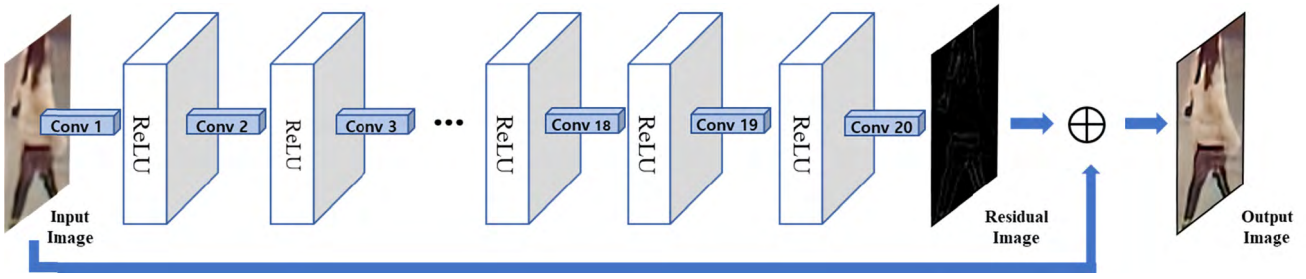


FIGURE 6. Architecture of VDSR.

C. MULTIMODAL GENDER RECOGNITION WITH DEEP RESIDUAL NETWORK

The images obtained through IRCNN and VDSR are input to ResNet-101 and used for gender recognition. In this study, 1:2.27 (197 × 447 pixels) images were used as input to ResNet-101, considering the body ratio of people, which are not square images (224 × 224 pixels) used in the traditional ResNet-101. As shown in Figure 7, when square images are used, much information, such as body shapes and ratios that represent the differences between men and women, is lost and the recognition performance decreases consequently. For the architecture of ResNet-101, a residual learning method of the shortcut structure is used as shown in Figure 8 and Table 4 [51]. This has the problem that more information is lost when the CNN is deeper. To solve this problem, ResNet-101 has a shortcut structure to prevent information loss by convolutional filters and to maintain information to solve these problems. Furthermore, Conv2 to Conv5 are bottleneck

TABLE 3. Detail descriptions of VDSR (N∗ denotes numbers from 1 to 19).

Layer type	Number of filters	Size of feature map (width × height × channel)	Size of kernel (width × height)	Number of strides	Number of paddings
Input layer [image]		W × H × 3			
Conv N* (N^{th} convolutional layer)	64	W × H × 64	3×3	1×1	1×1
ReLU N*		W × H × 64			
Conv 20	3	W × H × 3	3×3	1×1	1×1

FIGURE 7. Comparison of square images and images considering the body ratio of people. The human body ratio was considered for the above 197 × 447 images and the bottom 224 × 224 images are square image.

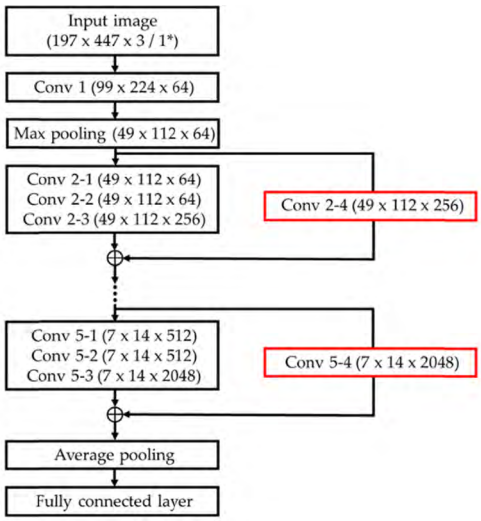


FIGURE 8. Architecture of ResNet-101. The red boxes indicate shortcuts (197 × 447 × 3 / 1∗ of input image indicates visible-light image (197 × 447 × 3) and IR image (197 x 447 x 1∗), respectively).

structures of 1 × 1, 3 × 3, and 1 × 1 convolutional filters, respectively. The purpose of the first 1 × 1 convolutional filter is to reduce the dimension as in the Inception structure of GoogleNet [46]. After reducing the dimension, it performs the 3 × 3 convolution, and then finally expands the dimension again through the 1 × 1 convolutional filter. Thus, the bottleneck structure reduces the computation amount through channel reduction by using the 1 × 1 convolutional filter.

In this study, we train visible-light and IR image, respectively, by the train from scratch method. The reason for using the two images, visible-light and IR images, for gender recognition is as follows. Most of the existing studies [4], [5], [20], [21], [27]–[29], [31], and [32] only used visible-light images. The visible-light images have the advantage of extracting detailed information of people, but they have the disadvantage of low recognition performance owing to illumination and shadow changes, background, clothes, and accessories. Therefore, we improved performance by combining visible-light images, which capture detailed information, with IR images, which are less affected by illumination and shadow changes, background, clothes, and accessories. In addition, the scalability was enhanced by combining various IR images for the IR cameras.

D. GENDER RECOGNITION BASED ON SCORE FUSION

Score fusion is performed based on the two score values (S_{vis} and S_{ir} of Equations (3) and (4)) that have passed through the fully connected layer of the ResNet-101 that uses the visible-light image as input and the ResNet-101 that uses IR image as input. Each score value before the score fusion is normalized by the min-max scaling method to have the range of 0 to 1. As shown in Equations (3) and (4), the score fusion was performed by applying the weighted sum and the weighted product rule, respectively.

Here, W is the optimal weight value determined through experiments.

$$\mathrm{WS} = \mathrm{W}S_{vis} + (1 - \mathrm{W})\, S_{ir} \tag{3}$$

$$\mathrm{WP} = S_{vis}^{\mathrm{W}} \cdot S_{ir}^{(1-\mathrm{W})} \tag{4}$$

The fused score determined thus (WS and WP) becomes the final score, which is used in gender recognition. In the gender recognition, gender is determined as male if the final score is greater than the threshold and as female if it is smaller than the threshold. At this moment, two error rates are generated: Type I error, which is the error rate that misclassifies male image into female image, and Type II error, which is the error rate that misclassifies female image with male image. In general, Type I error and Type II error have a trade-off relationship. Type II error decreases when Type I error increases, and Type II error increases when Type I error decreases. The error rate when Type I and Type II errors become equal is called the equal error rate (EER). In this

TABLE 4. Detailed descriptions of ResNet-101 (197 x 447 x 3 (197 x 447 x 1∗) of Input layer respectively indicates visible-light image (197 × 447 × 3) and IR image (197 × 447 × 1∗). The last convolution layer of each Conv (Conv 2-4∗, Conv 3-4∗, Conv 4-4∗, Conv 5-4∗) indicates a shortcut. 2 × 2 (1 × 1∗) indicates 2 × 2 only for the first iteration 2 × 2, and 1 × 1 for the other iterations).

Layer type		Number of filters	Size of feature map (width × height × channel)	Size of kernel (width × height)	Number of strides	Number of paddings	Number of iterations
Input layer [image]			197 × 447 × 3 (197 × 447 × 1*)				
Conv 1		64	99 × 224 × 64	7 × 7	2 × 2	3 × 3	1
Max pooling		1	49 × 112 × 64	3 × 3	2 × 2	0	1
Conv 2	Conv 2-1	64	49 × 112 × 64	1 × 1	1 × 1	0	3
	Conv 2-2	64	49 × 112 × 64	3 × 3	1 × 1	1 × 1	
	Conv 2-3	256	49 × 112 × 256	1 × 1	1 × 1	0	
	Conv 2-4*	256	49 × 112 × 256	1 × 1	1 × 1	0	
Conv 3	Conv 3-1	128	25 × 56 × 128	1 × 1	2 × 2 (1 × 1*)	0	4
	Conv 3-2	128	25 × 56 × 128	3 × 3	1 × 1	1 × 1	
	Conv 3-3	512	25 × 56 × 512	1 × 1	1 × 1	0	
	Conv 3-4*	512	25 × 56 × 512	1 × 1	2 × 2	0	
Conv 4	Conv 4-1	256	13 × 28 × 256	1 × 1	2 × 2 (1 × 1*)	0	23
	Conv 4-2	256	13 × 28 × 256	3 × 3	1 × 1	1 × 1	
	Conv 4-3	1024	13 × 28 × 1024	1 × 1	1 × 1	0	
	Conv 4-4*	1024	13 × 28 × 1024	1 × 1	2 × 2	0	
Conv 5	Conv 5-1	512	7 × 14 × 512	1 × 1	2 × 2 (1 × 1*)	0	3
	Conv 5-2	512	7 × 14 × 512	3 × 3	1 × 1	1 × 1	
	Conv 5-3	2048	7 × 14 × 2048	1 × 1	1 × 1	0	
	Conv 5-4*	2048	7 × 14 × 2048	1 × 1	2 × 2	0	
Average pooling		1	1 × 1 × 2048	7 × 7	1 × 1	0	1
Fully connected layer			2				1
Softmax layer			2				1

study, the threshold at the point where this ERR is obtained was used as the threshold for male and female determination.

V. EXPERIMENTAL RESULTS

A. EXPERIMENTAL DATABASE AND ENVIRONMENT

We perform gender recognition after two-step image reconstruction using the SYSU-MM01 database [7], which is an open database, as the first experimental data. For the SYSU-MM01 database, images were captured with six cameras in total: four visible-light cameras and two NIR cameras in the indoor and outdoor environments as shown in Figure 9. The visible-light images were captured during the day or afternoon, and the NIR images were captured in the evening or at night. Thus, the poses are different between the two types of images. Furthermore, as shown in the 2nd and 3rd column images in Figure 9, even though the people are of the same class, the clothes or accessories such as bag can be different between some of the visible-light and NIR images. The SYSU-MM01 database is composed of 491 people classes, 287,628 visible-light images, and 15,792 NIR images. In this study, 490 classes in total were used in the experiments after excluding classes that are different between the visible-light and NIR images. For the score fusion described in Section IV-D, the number of visible-light images must be the same as the number of NIR images. However, the number of visible-light images was much larger than the number

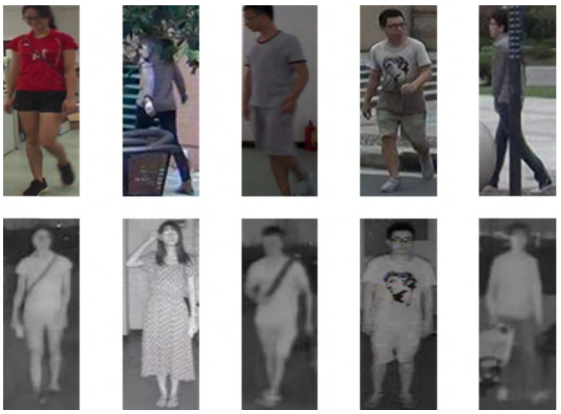


FIGURE 9. Example images of the SYSU-MM01 database. The visible-light images are at the top and the NIR images are at the bottom. Each column shows the images of the same class. The 1st and 2nd column images show females, and the other images show males.

TABLE 5. Descriptions of visible-light images in SYSU-MM01 database (unit: images).

Categories	Subset 1	Subset 2
Number of people (male / female)	245 (137 / 108)	245 (138 / 107)
Number of images (male / female)	7,724 (4,334 / 3,390)	7,771 (4,398 / 3,373)
Number of augmented images (male / female)	227,944 (112,684 / 115,260)	229,030 (114,348 / 114,682)

of NIR images. Thus, high-resolution images captured at a short distance for which gender recognition is relatively easy were excluded from the experiments. The experiments were conducted through two-fold cross validation, for which the entire database was divided into two subsets consisting of 245 classes that are different from each other (open world setting). Furthermore, to increase the number of images for training, data augmentation was performed by applying horizontal flipping, image shifting, and cropping. To prevent bias toward male or female data during training, we set the numbers of augmented female data to be different from those of male data in the case of image shifting. This data augmentation was performed only for the training data, and the testing was performed with the original data. Table 5 describes the numbers of original and augmented images in the two subsets of the SYSU-MM01 database used in this experiment and the total number of original and augmented images. Table 5 only shows the data for visible-light images. The values for the NIR images are the same as those in Table 5.

The algorithm proposed in this study was implemented by using MatConvNet (version 1.0-beta25) [55], Caffe framework (version 1.0.0) [47], Microsoft Visual Studio 2013, and OpenCV (version 3.4.5) [56]. The experiments were conducted using Intel®Core^{TM}i7-7700 CPU @ 3.6 GHz (4 cores) with 24 GB of main memory, and NVIDIA GeForce GTX 1070 Ti (2432 compute unified device architecture (CUDA) cores) with graphics memory of 8 GB (NVIDIA, Santa Clara, CA, USA) [52].

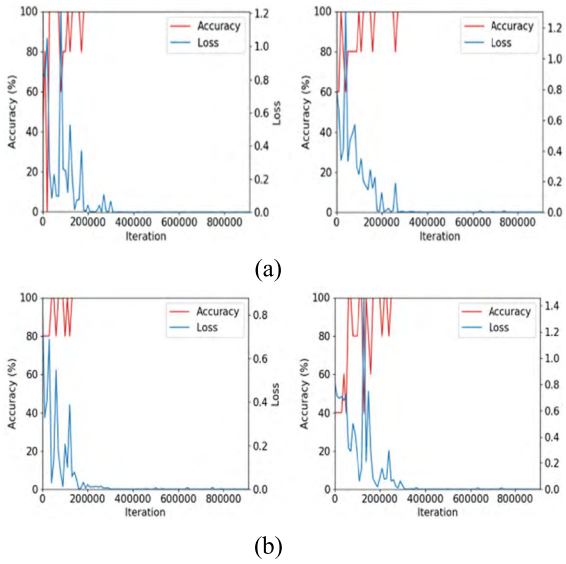


FIGURE 10. Examples of training loss and accuracy graphs with SYSU-MM01 database. Left and right figures are the training graphs with subset 1 and subset 2 of Table 5, respectively. Training graphs with (a) visible-light images and (b) NIR images, respectively.

B. TRAINING

Stochastic gradient descent (SGD) [53] was used to train ResNet-101 with the training data obtained by data augmentation. A characteristic of the SGD is that training was performed in mini-batch size units. The number of iterations is calculated by "number of training data / mini-batch size," and the number of iterations determined by this calculation is defined as 1 epoch. For visible-light images with two-step image reconstruction, 0.9 was used for momentum, 0.0001 for weight decay, and 0.1 for the initial learning rate. For NIR images, 0.9 was used for momentum, 0.0001 for weight decay, and 0.1 for the initial learning rate.

At this time, optimization was performed by using the step policy, which multiplies the gamma value each time after a fixed number of iterations, as the learning rate policy. Training was performed for 20 epochs and the step size was set as approximately 3 epochs. Figure 10 shows the graphs of training loss and accuracy. As the training iteration increased, the training loss converged to 0 and the training accuracy to 100%. This indicates that the training of the ResNet-101 for gender recognition used in this study was performed successfully. The reason why the training accuracies were shown by the unit of 20% in Figure 10 is because the training was performed with the mini-batch size of 5 in our experiments. In details, SGD method performs the training based on mini-batch instead of whole training data [53]. Because the numbers of mini-batch is 5, the training accuracies are shown by the unit of 20%, that is, one of the five values such as 0% ((0/5)×100), 20% ((1/5)×100), 40% ((2/5)×100), 60% ((3/5)×100), 80% ((4/5)×100), and 100% ((5/5)×100).

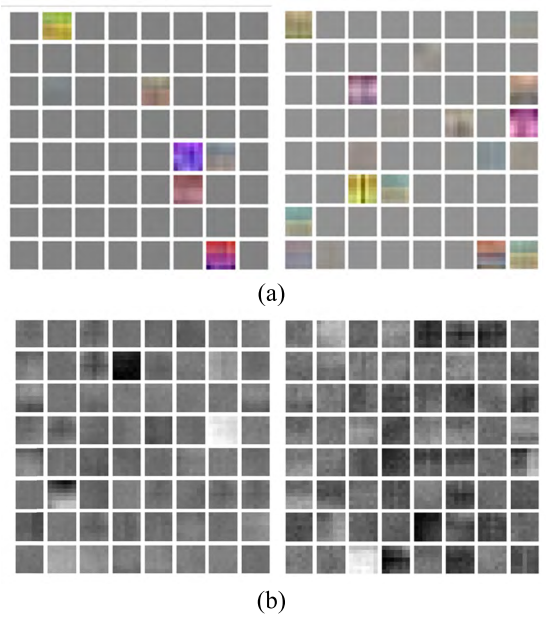


FIGURE 11. Example of trained filter images. Left and right figures are the filter image with subset 1 and subset 2 of Table 5, respectively. Trained filter images with (a) visible-light image and (b) NIR image.

Figure 11 shows examples of trained filter images. These filters are used in Conv1 in Table 4. As in Table 4, 64 filters of size 7 × 7 are shown. In this paper, the 7 × 7 size was enlarged for visibility. Also, as in the input layer of Table 4, visible-light images have 3 channels, so the trained filters with visible-light images have 3 channels whereas NIR images have 1 channel, so the trained filters with NIR images have 1 channel. As shown in Figure 11, the appearances of trained filters with visible-light image and NIR image are very different. As described in Section I, visible-light images include much variations of illumination, background, color, clothes, and accessories than those in NIR images. These variations are not useful features for the network to improve gender recognition accuracy. Therefore, the amount of distinctive textures trained by CNN, which is meaningful in the gender recognition, is less in the trained filters of Figure 11 (a) compared to that in the trained filters with NIR images of Figure 11 (b). In addition, the optimal kernel size in CNN was experimentally determined with training data. That is, the useless features affected by appearance variation can be removed through the filters obtained by training for high accuracy of gender recognition.

C. TESTING OF PROPOSED METHOD WITH SYSU-MM01 DATABASE

The EER and receiver operating characteristic (ROC) curve was used for comparison and analysis of the testing accuracy of the proposed method. As this study performed score fusion using two image types, visible-light and NIR images, each performance was compared first and then the score fusion performance was compared.

TABLE 6. Comparisons of gender recognition accuracies with visible-light images.

Methods	Two-fold cross validation	EER (%)	
		1st and 2nd fold	Average
Original image	1st fold	14.73	14.39
	2nd fold	14.05	
Original image + IRCNN	1st fold	15.41	15.21
	2nd fold	15.01	
Original image + DCSCN	1st fold	14.70	14.37
	2nd fold	14.04	
Original image + VDSR	1st fold	14.26	13.97
	2nd fold	13.68	
Original image + IRCNN + VDSR (proposed method)	1st fold	13.78	**13.63**
	2nd fold	13.48	

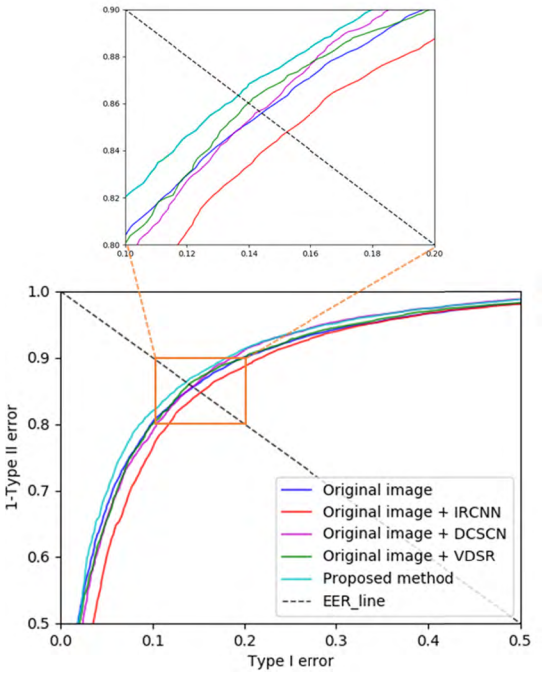


FIGURE 12. ROC curves of gender recognition accuracy with visible-light images.

1) COMPARISONS OF RECOGNITION ACCURACIES WITH SINGLE IMAGE

Table 6 and Figure 12 show performance measurements for various methods using visible-light images. IRCNN was used for denoising as mentioned in Section IV-B-2. Deep CNN with residual net, skip connection and network in network (DCSCN) [48], and VDSR were used for image SR reconstruction. As shown in Table 6 and Figure 12, the results of performance comparison between DCSCN and VDSR showed that the VDSR had a higher performance. Thus, we adopt VDSR as the SR reconstruction method and we measure the gender recognition performance by combining

TABLE 7. Comparisons of gender recognition accuracies with NIR images.

Methods	Two-fold cross validation	EER (%)	
		1st and 2nd fold	Average
Original image (proposed method)	1st fold	9.25	**9.02**
	2nd fold	8.79	
Original image + IRCNN	1st fold	9.25	9.60
	2nd fold	9.95	
Original image + DCSCN	1st fold	17.57	14.33
	2nd fold	11.09	
Original image + VDSR	1st fold	8.99	9.67
	2nd fold	10.35	
Original image + IRCNN + VDSR	1st fold	10.83	11.10
	2nd fold	11.37	

the IRCNN and VDSR. As shown in Table 6, the performance only by VDSR was lower than that by both IRCNN and VDSR. In addition, as shown in Table 6 and Figure 12, the performance decreased when only the IRCNN was applied to the original image. The reason for this appears to be that, in the case of the IRCNN, the residual image is subtracted from the original image.

However, the original image is already a low-resolution image with a small amount of information, and the performance decreased because the information was subtracted. In the case of the DCSCN and VDSR, although the performance increased compared with that of the original image, the differences is small. Finally, after removing unnecessary noises from the original image, image reconstruction was performed through the VDSR, which has a higher performance than the DCSCN. Consequently, the highest recognition performance was obtained.

In the next experiment, performance was measured for various methods using the NIR image as shown in Table 7 and Figure 13. Similarly, the VDSR was adopted as the SR reconstruction method because it showed a higher performance than the DCSCN, and the performance of the images was measured by combining the IRCNN and VDSR. As shown in the experimental results in Table 7 and Figure 13, the gender recognition performance was the highest when the original images were used without using the IRCNN, DCSCN, and VDSR. This is because the NIR image is less affected by other factors such as background because it is composed of a single channel unlike the visible-light image. Thus, subtracting residual information from the image (IRCNN) or adding it to the image (DCSCN, VDSR) resulted in lowering the performance on the contrary.

2) MEASURING RECOGNITION ACCURACIES BASED ON SCORE LEVEL FUSION AND COMPARISONS WITH PREVIOUS METHODS

Table 8 and Figure 14 compare the performances of the case of using individual visible-light and NIR images and the case

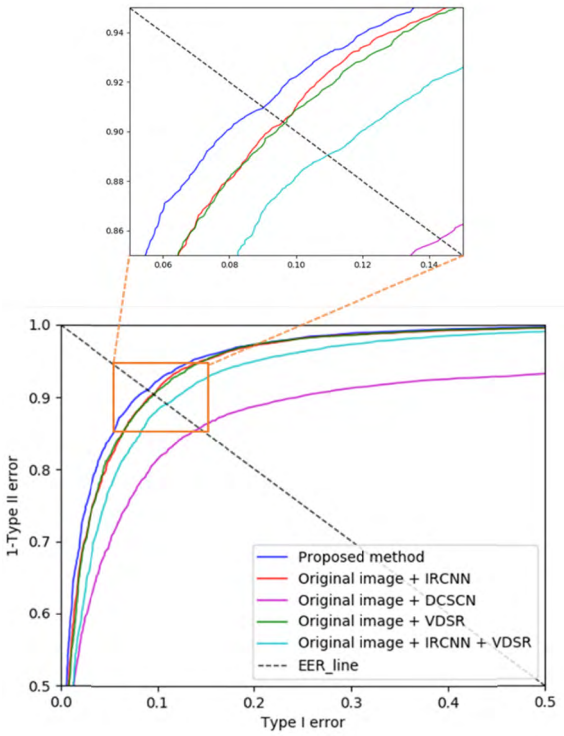


FIGURE 13. ROC curves of gender recognition accuracy with NIR images.

TABLE 8. Comparisons of gender recognition accuracies by single image and multimodal images using score fusion.

Methods	Two-fold cross validation	EER (%)	
		1st and 2nd fold	Average
Only using visible-light image (original image + IRCNN + VDSR)	1st fold	13.78	13.63
	2nd fold	13.48	
Only using NIR image (original image)	1st fold	9.25	9.02
	2nd fold	8.79	
Weighted product	1st fold	5.01	5.28
	2nd fold	5.55	
Weighted sum (proposed method)	1st fold	5.03	**5.27**
	2nd fold	5.51	

of score fusion using the weighted product and weighted sum methods described in Section IV-D. The experimental results show the highest performance for the fusion using the weighted sum.

Figures 15 and 16 show the successful and failed results of gender recognition, respectively. In Figure 15, successful results were obtained even for images in which parts of the faces and bodies were occluded, making it difficult to recognize the gender. In Figure 16, recognition failed because

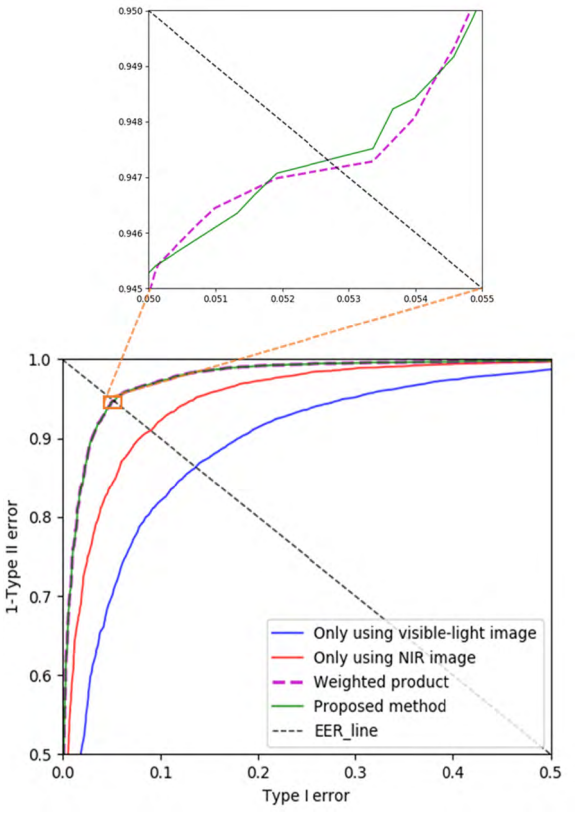


FIGURE 14. ROC curves of gender recognition accuracy by single image and multimodal images using score fusion.

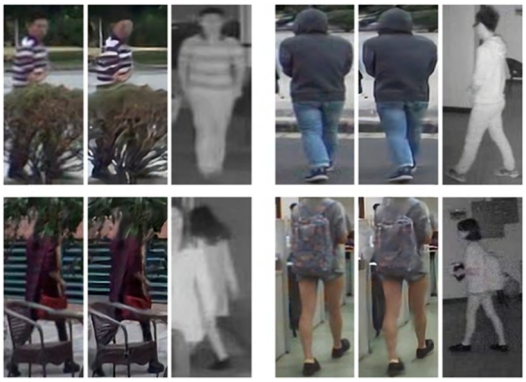


FIGURE 15. Examples of successful gender recognition. From left to right, the original visible-light image, reconstructed visible-light image, and original NIR image, respectively. The top images are male images and the bottom images are female images.

it was difficult to recognize the gender from the image itself or the body was hidden by some objects.

Table 9 compares the recognition performance between the proposed method and existing studies [4], [28], [33]–[35]. For a fair comparison, cases using a single visible-light image, a single NIR image, and score fusion were compared separately. The experimental results showed that the proposed method has a higher performance than the existing methods. The weighted HOG method [34] showed a lower performance than the HOG method [4], [33]. This is because, for the dataset of SYSU-MM01, as calibration has not been

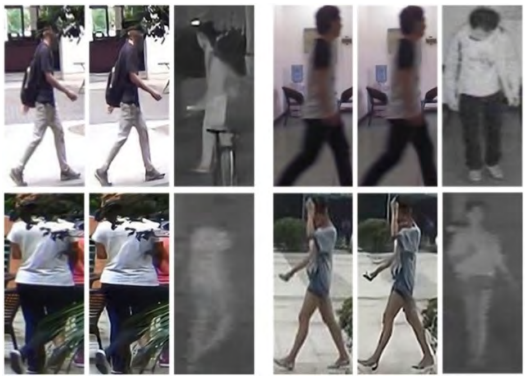


FIGURE 16. Examples of failed gender recognition. From left to right, the original visible-light image, reconstructed visible-light image, and original NIR image, respectively. The top images are of males, but were recognized as female. The bottom images are of females, but were recognized as males.

TABLE 9. Comparisons of gender recognition accuracies by our method with previous methods (unit: %).

Methods	Using single visible-light images	Using single NIR images	Score fusion (weighted sum)
HOG [4, 33]	21.51	19.19	18.51
Weighted HOG [34]	28.84	25.46	23.90
AlexNet [28, 35]	32.72	38.99	24.53
Proposed method	13.63	9.02	**5.27**

performed between the visible-light camera and NIR camera, the weight obtained from the NIR image is meaningless in the visible-light image; thus, the performance was lower. Furthermore, the study in [34] used thermal images, but SYSU-MM01 consists of NIR image data that have no clear distinction between background and people regions, and this is also a cause of the decreased performance.

The performance of the AlexNet-based method was the lowest, because SYSU-MM01 database consists of many images that have the part of human body occluded and large pose variations, which caused that shallow AlexNet not to be properly trained.

D. TESTGING OF PROPOSED METHOD WITH DBGENDER-DB2 DATABASE

To verify whether the proposed method also works in various combinations of heterogeneous camera images, the DBGender-DB2 database built by us [8] was used in the next experiment. The images in the DBGender-DB2 database were captured using one visible-light camera and one thermal camera in an outdoor environment. The SYSU-MM01 database contained some short-distance images because they were captured in an indoor and outdoor environment. However, the DBGender-DB2 database consists

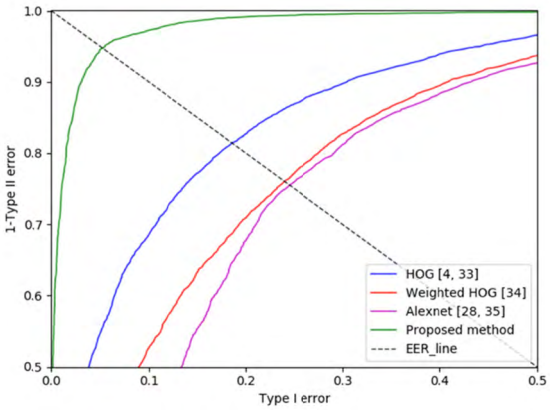


FIGURE 17. ROC curves of gender recognition accuracy by the proposed method and previous methods.

FIGURE 18. Example images of the DBGender-DB2 database. The visible-light images are at the top and thermal images are at the bottom. The images in the 1st and 2nd columns are female, and the other images are male.

of low-resolution images only, which were captured in outdoor environments as shown in Figure 18. Furthermore, as shown in Figure 18, visible-light and thermal images were captured simultaneously for the DBGender-DB2, which consists of images of the same poses.

The DBGender-DB2 database include 412 people classes, 4,120 visible-light images, and 4,120 thermal images in total. As the total number of images was small, five-fold cross validation was performed instead of two-fold cross validation. Table 10 describes the numbers of original and augmented images for males and females in five subsets of the DBGender-DB2 database used in the experiments. When five-fold cross validation was performed for the DBGender-DB2 database, the numbers of training sets and test sets of each folds is same. Thus, only the case of one fold is presented in Table 10. The other four folds are performed with the same number as well. Furthermore, as the number of visible-light images is the same as the number of thermal images, Table 10 only presents the visible-light images. In the case of the DBGender-DB2 database, the number of augmented images was smaller than that of the SYSU-MM01 database. Thus, we perform training for 60 epochs and the other

TABLE 10. Descriptions of visible-light images in DBGender-DB2 database (unit: images).

Categories	Training set	Test set
Number of people (male / female)	331 (204 / 127)	81 (50 / 31)
Number of images (male / female)	74,820 (36,720 / 38,100)	1,620 (1,000 / 620)
Number of augmented images (male / female)	149,640 (73,440 / 76,200)	No augmentation

TABLE 11. Comparisons of gender recognition accuracies with visible-light images.

Methods	Five-fold cross validation	EER (%)	
		1^{st}–5^{th} fold	Average
Original image	1^{st} fold	15.79	18.56
	2^{nd} fold	21.33	
	3^{rd} fold	22.21	
	4^{th} fold	18.02	
	5^{th} fold	15.43	
Original image + IRCNN + VDSR (proposed method)	1^{st} fold	12.66	**14.62**
	2^{nd} fold	16.65	
	3^{rd} fold	19.25	
	4^{th} fold	12.46	
	5^{th} fold	12.10	

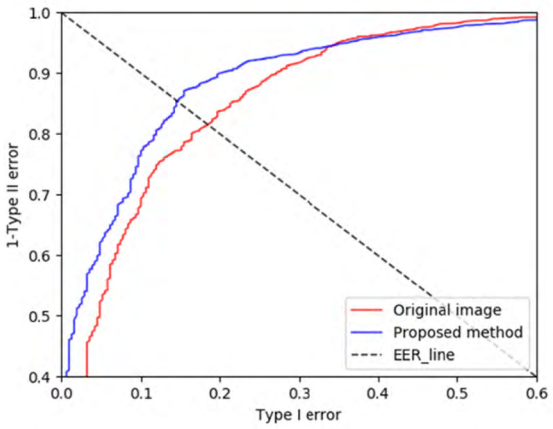


FIGURE 19. ROC curves of gender recognition accuracy with visible-light images.

parameter values were set the same as in the case of the SYSU-MM01 database.

For the experiment of the DBGender-DB2 database, the performance results were compared between the original image and the original image + IRCNN + VDSR, which showed good performance among the experiments using the SYSU-MM01 database described in Section V-C-1. Furthermore, the performance was compared between the proposed method and existing studies. Tables 11 and 12, and

TABLE 12. Comparisons of gender recognition accuracies with thermal images.

Methods	Five-fold cross validation	EER (%)	
		1st–5th fold	Average
Original image (proposed method)	1st fold	14.46	**13.87**
	2nd fold	9.02	
	3rd fold	16.16	
	4th fold	10.98	
	5th fold	18.75	
Original image + IRCNN + VDSR	1st fold	21.97	17.61
	2nd fold	15.08	
	3rd fold	15.54	
	4th fold	16.79	
	5th fold	18.65	

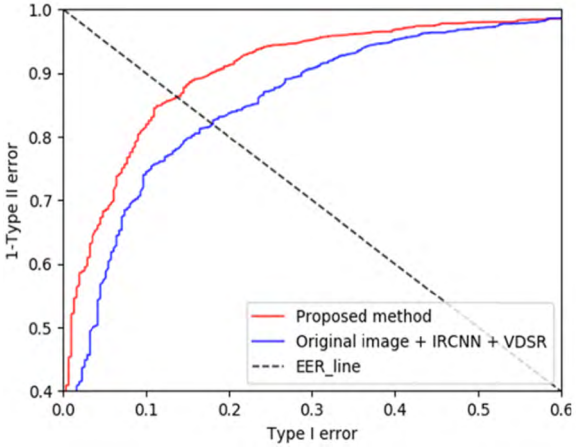


FIGURE 20. ROC curves of gender recognition accuracy with thermal images.

Figures 19 and 20 show the recognition performance for visible-light and thermal images, respectively. In the DBGender-DB2 database as well, applying both IRCNN and VDSR showed better performance for visible-light images compared with that of the original image. However, for thermal images, applying both IRCNN and VDSR showed lower performance compared with that of the original image. The reason for this is the same as in the case of SYSU-MM01 database. In the thermal image, the background and people are already distinguished, and subtracting the residual image (IRCNN) from or adding it (VDSR) to the original image distorts the shape of people, thus lowering the performance.

Table 13 and Figure 21 compare the performance between using individual visible-light and NIR images and using score fusion by weighted sum with the weighted product as described in Section IV-D. The experiment results showed that fusion using the weighted sum had the best performance as in the SYSU-MM01 database. Figures 22 and 23 show the successful and failed cases of gender recognition.

TABLE 13. Comparisons of gender recognition accuracies by single image and multimodal images using score fusion.

Methods	Five-fold cross validation	EER (%)	
		1st ~ 5th fold	Average
Only using visible-light image (original image + IRCNN + VDSR)	1st fold	12.66	14.62
	2nd fold	16.65	
	3rd fold	19.25	
	4th fold	12.46	
	5th fold	12.10	
Only using thermal image (original image)	1st fold	14.46	13.87
	2nd fold	9.02	
	3rd fold	16.16	
	4th fold	10.98	
	5th fold	18.75	
Weighted product	1st fold	13.75	12.05
	2nd fold	9.97	
	3rd fold	12.97	
	4th fold	8.79	
	5th fold	14.75	
Weighted sum (proposed method)	1st fold	12.95	**10.98**
	2nd fold	8.03	
	3rd fold	12.07	
	4th fold	8.39	
	5th fold	13.47	

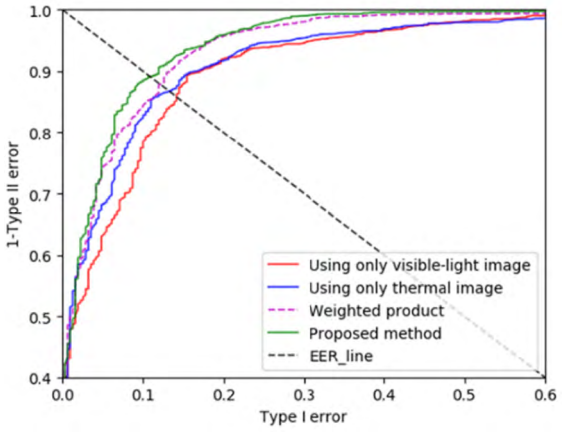


FIGURE 21. ROC curves of gender recognition accuracy by single image and multimodal images using score fusion.

In Figure 22, the gender recognition was successful even though the gender of the images was difficult to recognize. In Figure 23, the gender was not recognized because the image resolution was too low to distinguish gender or distortion occurred in the image reconstruction process.

Table 14 compares the recognition performance between the proposed method and existing studies [4], [28], [33]–[35]. For a fair comparison, cases using single visible-light image, single NIR image, and score fusion were compared. The experimental results show that the proposed method has a

FIGURE 22. Successful cases of gender recognition. From left to right, the original visible-light image, reconstructed visible-light image, and original thermal image, respectively. The top images are of males, and the bottom images are of females.

FIGURE 23. Failed cases of gender recognition. From left to right, the original visible-light image, reconstructed visible-light image, and original thermal image, respectively. The top images are male but recognized as female; the bottom images are female but recognized as male.

TABLE 14. Comparisons of gender recognition accuracies by our method with previous methods (unit: %).

Methods	Using single visible-light images	Using single NIR images	Score fusion
HOG [4, 33]	17.82	20.46	16.28
Weighted HOG [34]	15.22	18.26	13.06
AlexNet [28, 35]	17.06	16.11	11.71
Proposed method	14.62	13.87	**10.98**

better performance than the existing methods. The images in the DBGender-DB2 database were acquired under the condition that the visible-light and thermal cameras are calibrated to each other. Furthermore, for thermal images, which have clear distinction between background and original images, weighted HOG showed a higher performance than that in the SYSU-MM01 database. The proposed method showed a higher performance for single images as well as for score fusion than previous methods.

TABLE 15. Average processing time of proposed method (unit: ms).

	IRCNN	VDSR	ResNet and score fusion	Total
Desktop computer	2.08	4.23	24.69	31
Jetson TX2 embedded system	3.94	4.54	90.78	99.26

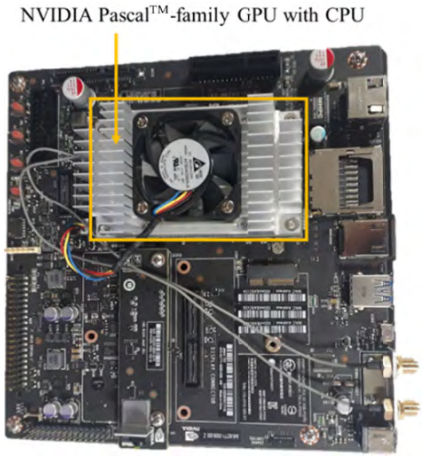


FIGURE 24. Jetson TX2 embedded system.

E. PROCESSING TIME

Table 15 shows the measurements of average processing time per image of the proposed method in a desktop environment as described in Section V-A. As shown in this table, the average processing time per image was 31 ms, which suggests that the proposed method can process images at the speed of approximately 32 frames per sec. In the next experiment, the processing time was measured in the Jetson TX2 embedded system [54], which is often used for on-board deep learning processing in autonomous vehicles as shown in Figure 24. Jetson TX2 has NVIDIA Pascal™-family GPU (256 CUDA cores), having 8 GB of memory shared between the central processing unit (CPU) and GPU, and 59.7 GB/s of memory bandwidth; it uses less than 7.5 W of power. As shown in Table 15, the average processing time per image was 99.26 ms, which suggests that the proposed method can process images in this environment at the speed of approximately 10 frames per sec. The Jetson TX2 embedded system required a longer processing time than the desktop computer owing to limited computing resources. However, this result confirms that the proposed method is also applicable to embedded systems with limited computing resources.

F. ANALYSIS OF FEATURE MAP

In general, as the depth of the convolutional layer increases, the size (width and height) of the feature map becomes smaller, but the number of channels of the feature

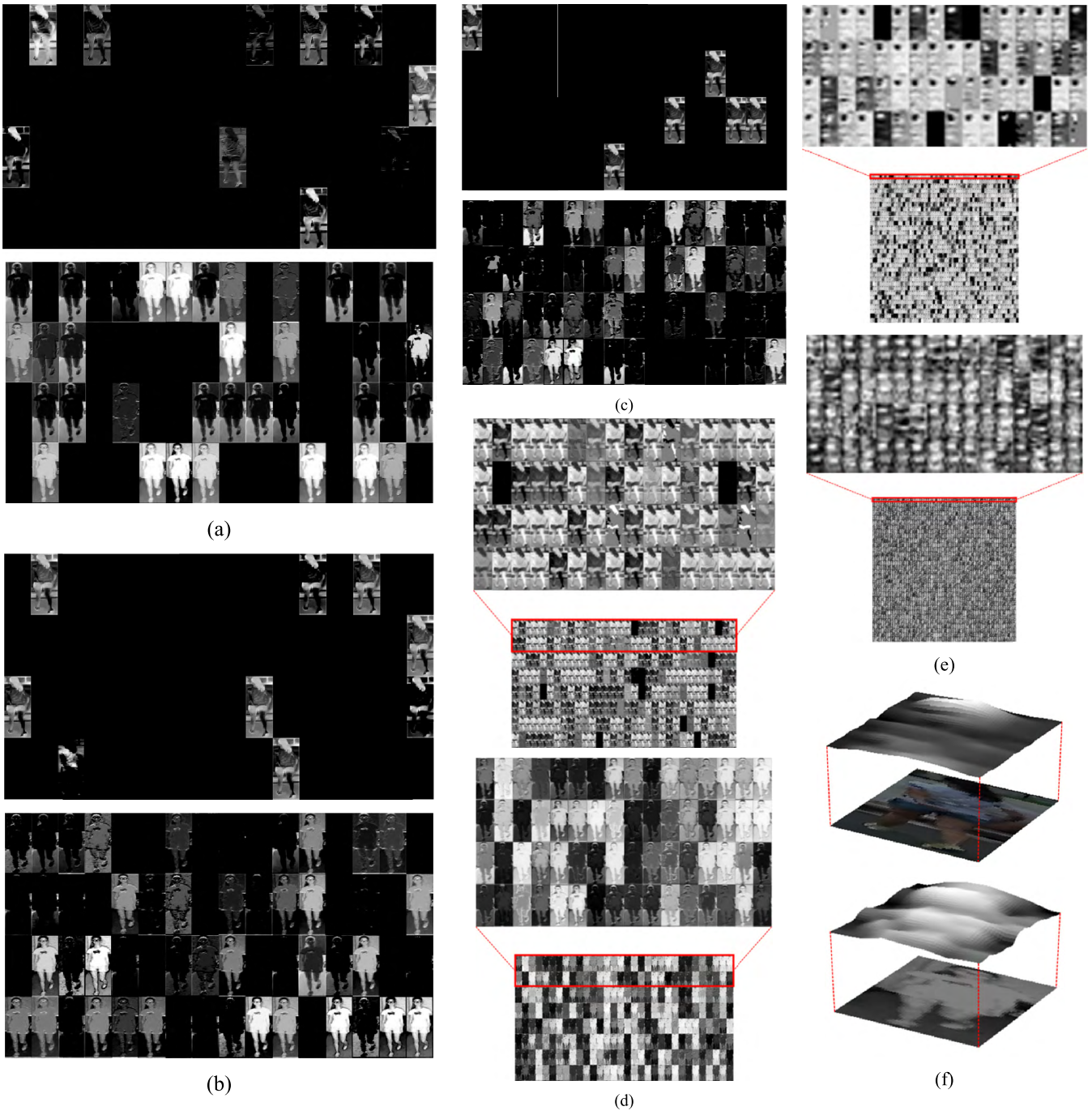


FIGURE 25. Examples of feature maps extracted from each layer for the input images. In (a)–(f), upper and lower figures are obtained with visible-light and NIR images of SYSU-MM01 database, respectively. Feature maps from (a) conv 1 of Table 4, (b,c,d) the first, second, and third iterations of residual blocks in conv 2 of Table 4, respectively, and (e) the last residual blocks in conv 5 of Table 4. (f) Three-dimensional feature map image obtained by averaging all the feature map values of (e).

map increases. Furthermore, the number of applied filters is smaller in the layer closer to the input with large images, and the number of applied filters is larger in the deeper layer that is farther from the input. In this sub-section, the feature maps obtained from ResNet-101 are analyzed as shown in Figure 25.

As shown in Figure 25, the deeper the layer of the CNN, the higher is the depth of feature maps. Figure 25 (a) shows the feature map obtained through Conv 1 of Table 4. Figures 25 (b), (c), and (d) show feature maps obtained from the first, second, and third iterations of residual blocks in Conv 2 of Table 4, respectively. Figures 25 (e) shows the feature map of the last residual blocks in Conv 5 of Table 4. As mentioned above, Figure 25 confirms that, as the layer becomes deeper, the depth of the feature map increases, and the features become more abstracted. Figure 25 (f) shows the

three-dimensional feature map image obtained by averaging the feature map values of all channels in Figure 25 (e). As shown in the upper image of Figure 25 (f) (visible-light image), the magnitudes are particularly large in the hair styles and little in the human body area because the variations on cloths, color, illuminations, and accessories, etc cause the lower meaningful features in the human body area. However, the magnitudes are large in the human body region as shown in the lower image of Figure 25 (f) (NIR image). As shown in the upper images of Figures 25 (a)-(e) (visible-light image), there are more images with all pixel values of 0 in the feature maps compared to those obtained from NIR images because of the larger variations of cloths, color, illuminations, background, and accessories, etc in visible-light images than those in NIR images as explained in Section V-B based on Figure 11. The degradation of gender recognition accuracy caused by the appearance variation in visible-light image can be lessened by fusing the two recognition scores from visible-light and NIR images in our research.

VI. CONCLUSION

In this study, we proposed a method that improve gender recognition performance for low-resolution human full-body images. Generally, the human full-body images captured at a distance in surveillance systems have poor image quality and low recognition performance owing to illumination variations, shadows, background, clothes, and accessories. To improve the gender recognition performance of these human full-body images, two types of CNN were used to remove the noise of visible-light images and enhance the image quality. Furthermore, in addition to visible-light images, IR images were used to reduce the effects of illumination variations, shadows, background, clothes, and accessories on gender recognition. In addition, the scalability of gender recognition using the combination of visible-light images and various IR images as well as the recognition performance were improved through comparative experiments using two open databases. A comparison with existing studies showed that the proposed method has a higher performance. Considering on-board application in the future, the processing speed of the proposed method was measured not only in a desktop computer, but also in a Jetson TX2 embedded system, thus verifying applicability to various platforms.

In future work, we will investigate methods to improve the performance further by combining optical and motion blurring restoration, SR reconstruction, and denoising methods for input images captured at a farther distance. In addition, we will investigate methods to improve recognition performance through image reconstruction based on a generative adversarial network.

REFERENCES

[1] S. Baluja and H. A. Rowley, "Boosting sex identification performance," *Int. J. Comput. Vis.*, vol. 71, no. 1, pp. 111–119, Jan. 2007.

[2] J. Mansanet, A. Albiol, and R. Paredes, "Local deep neural networks for gender recognition," *Pattern Recognit. Lett.*, vol. 70, pp. 80–86, Jan. 2016.

[3] C. Shan, "Learning local binary patterns for gender classification on real-world face images," *Pattern Recognit. Lett.*, vol. 33, no. 4, pp. 431–437, Mar. 2012.

[4] L. Cao, M. Dikmen, Y. Fu, and T. S. Huang, "Gender recognition from body," in *Proc. 16th ACM Int. Conf. Multimedia*, Oct. 2008, pp. 725–728.

[5] M. Collins, J. Zhang, P. Miller, and H. Wang, "Full body image feature representations for gender profiling," in *Proc. IEEE 12th Int. Conf. Comput. Vis. Workshops*, Kyoto, Japan, Sep./Oct. 2009, pp. 1235–1242.

[6] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 21-26 Jul. 2017, pp. 2790–2798.

[7] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai, "Rgb-infrared cross-modality person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, Oct. 2017, pp. 5380–5389.

[8] *Dongguk Body-based Gender Database (DBGender-DB2)*. Accessed: Jan. 2017. [Online]. Available: http://dm.dgu.edu/link.html

[9] C. B. Ng, Y. H. Tay, and B. M. Goi, "Vision-based human gender recognition: A survey," Apr. 2012, *arXiv:1204.1611*. [Online]. Available: https://arxiv.org/abs/1204.1611

[10] B. Moghaddam and M.-H. Yang, "Learning gender with support faces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 707–711, May 2002.

[11] Z. Sun, G. Bebis, X. Yuan, and S. J. Louis, "Genetic feature subset selection for gender classification: A comparison study," in *Proc. 6th IEEE Workshop Appl. Comput. Vis.*, Orlando, FL, USA, Dec. 2002, pp. 165–170.

[12] G. Shakhnarovich, P. A. Viola, and B. Moghaddam, "A unified learning framework for real time face detection and classification," in *Proc. 5th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Washington, DC, USA, May 2002, pp. 16–23.

[13] G. Levi and T. Hassncer, "Age and gender classification using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Boston, MA, USA, Jun. 2015, pp. 34–42.

[14] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019.

[15] L. Lee and W. E. L. Grimson, "Gait analysis for recognition and classification," in *Proc. 5th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Washington, DC, USA, May 2002, pp. 155–162.

[16] S. Yu, T. Tan, K. Huang, K. Jia, and X. Wu, "A study on gait-based gender classification," *IEEE Trans. Image Process.*, vol. 18, no. 8, pp. 1905–1910, Aug. 2009.

[17] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006.

[18] J. Lu, G. Wang, and P. Moulin, "Human identity and gender recognition from gait sequences with arbitrary walking directions," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 1, pp. 51–61, Jan. 2014.

[19] S. Yu, D. Tan, and T. Tan, "Modelling the effect of view angle variation on appearance-based gait recognition," in *Proc. Asian Conf. Comput. Vis.*, Hyderabad, India, Jan. 2006, pp. 807–816.

[20] L. Bourdev, S. Maji, and J. Malik, "Describing people: A poselet-based approach to attribute classification," in *Proc. Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 1543–1550.

[21] G. Guo, G. Mu, and Y. Fu, "Gender from body: A biologically-inspired approach with manifold learning," in *Proc. Asian Conf. Comput. Vis.*, Xi'an, China, Sep. 2009, pp. 236–245.

[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, Jun. 2005, pp. 886–893.

[23] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proc. 13th Int. Conf. Int. Conf. Mach. Learn.*, Bari, Italy, Jul. 1996, pp. 148–156.

[24] *Label Information of Sun Yat-sen University Multiple Modality Re-ID (SYSU-MM01) Database and CNN Models*. Accessed: May 2019. [Online]. Available: http://dm.dgu.edu/link.html

[25] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3D human pose annotations," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Kyoto, Japan, Sep./Oct. 2009, pp. 1365–1372.

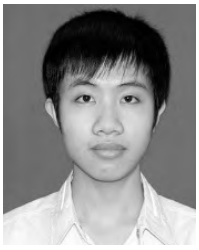
[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, Dec. 2012, pp. 1097–1105.

[27] C.-B. Ng, Y.-H. Tay, and B.-M. Goi, "A convolutional neural network for pedestrian gender recognition," in *Proc. Int. Symp. Neural Netw.*, Jul. 2013, pp. 558–564.
[28] G. Antipov, S.-A. Berrani, N. Ruchaud, and J.-L. Dugelay, "Learned vs. hand-crafted features for pedestrian gender recognition," in *Proc. 23rd ACM Int. Conf. Multimedia*, Brisbane, Australia, Oct. 2015, pp. 1263–1266.
[29] C.-B. Ng, Y.-H. Tay, and B.-M. Goi, "Pedestrian gender classification using combined global and local parts-based convolutional neural networks," *Pattern Anal. Appl.*, pp. 1–12, Jul. 2018. doi: 10.1007/s10044-018-0725-0.
[30] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent.*, San Diego, CA, USA, May 2015, pp. 1–14.
[31] L. Cai, J. Zhu, H. Zeng, J. Chen, C. Cai, and K.-K. Ma, "HOG-assisted deep feature learning for pedestrian gender recognition," *J. Franklin Inst.*, vol. 355, no. 4, pp. 1991–2008, Mar. 2018.
[32] M. Raza, M. Sharif, M. Yasmin, M. A. Khan, T. Saba, and S. L. Fernandes, "Appearance based pedestrians' gender recognition by employing stacked auto encoders in deep learning," *Future Gener. Comput. Syst.*, vol. 88, pp. 28–39, Nov. 2018.
[33] D. T. Nguyen and K. R. Park, "Body-based gender recognition using images from visible and thermal cameras," *Sensors*, vol. 16, no. 2, p. 156, 2016.
[34] D. T. Nguyen and K. R. Park, "Enhanced gender recognition system using an improved histogram of oriented gradient (HOG) feature from quality assessment of visible light and thermal images of the human body," *Sensors*, vol. 16, no. 7, p. 1134, 2016.
[35] D. T. Nguyen, K. W. Kim, H. G. Hong, J. H. Koo, M. C. Kim, and K. R. Park, "Gender recognition from human-body images using visible-light and thermal camera videos based on a convolutional neural network for image feature extraction," *Sensors*, vol. 17, no. 3, p. 637, 2017.
[36] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep CNN denoiser prior for image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2808–2817.
[37] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 1646–1654.
[38] J. Sun, Z. Xu, and H.-Y. Shum, "Image super-resolution using gradient profile prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Anchorage, AK, USA, Jun. 2008, pp. 1–8.
[39] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Washington, DC, USA, Jun./Jul. 2004, p. I.
[40] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Comput. Graph. Appl.*, vol. 22, no. 2, pp. 56–65, Mar./Apr. 2002.
[41] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proc. Int. Conf. Curves Surfaces*, Avignon, France, Jun. 2010, pp. 711–730.
[42] T. Peleg and M. Elad, "A statistical prediction model based on sparse representations for single image super-resolution," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2569–2582, Jun. 2014.
[43] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Kyoto, Japan, Sep./Oct. 2009, pp. 349–356.
[44] S. Gu, N. Sang, and F. Ma, "Fast image super resolution via local regression," in *Proc. 21st Int. Conf. Pattern Recognit.*, Tsukuba, Japan, Nov. 2012, pp. 3128–3131.
[45] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.
[46] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 1–9.
[47] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, Orlando, FL, USA, Nov. 2014, pp. 675–678.
[48] J. Yamanaka, S. Kuwashima, and T. Kurita, "Fast and accurate image super resolution by deep CNN with skip connection and network in network," in *Proc. Int. Conf. Neural Inf. Process.*, Guangzhou, China, Nov. 2017, pp. 217–225.
[49] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: A technical overview," *IEEE Signal Process. Mag.*, vol. 20, no. 3, pp. 21–36, May 2003.
[50] K. Nasrollahi and T. B. Moeslund, "Super-resolution: A comprehensive survey," *Mach. Vis. Appl.*, vol. 25, no. 6, pp. 1423–1468, Aug. 2014.
[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.
[52] *GEFORCE GTX 1070 Ti*. Accessed: May 2019. [Online]. Available: https://www.nvidia.com/en-us/geforce/products/10series/geforce-gtx-1070-ti/
[53] L. Bottou, "Stochastic gradient descent tricks," in *Neural Networks: Tricks of the Trade*. Berlin, Germany: Springer, 2012, pp. 421–436.
[54] *Jetson TX2 Module*. Accessed: Feb. 2019. [Online]. Available: https://www.nvidia.com/en-us/autonomous-machines/embedded-systems-dev-kits-modules/
[55] A. Vedaldi and K. Lenc, "Matconvnet: Convolutional neural networks for MATLAB," in *Proc. 23rd ACM Int. Conf. Multimedia*, Brisbane, Australia, Oct. 2015, pp. 689–692.
[56] *OpenCV Library*. Accessed: Dec. 2018. [Online]. Available: https://opencv.org/releases/
[57] X. Lan, M. Ye, R. Shao, B. Zhong, P. C. Yuen, and H. Zhou, "Learning modality-consistency feature templates: A robust RGB-infrared tracking system," *IEEE Trans. Ind. Electron.* to be published.
[58] X. Lan, M. Ye, R. Shao, B. Zhong, D. K. Jain, and H. Zhou, "Online non-negative multi-modality feature template learning for RGB-assisted infrared tracking," *IEEE Access*, vol. 7, pp. 67761–67771, 2019.
[59] X. Lan, M. Ye, S. Zhang, H. Zhou, and P. C. Yuen, "Modality-correlation-aware sparse representation for RGB-infrared object tracking," *Pattern Recognit. Lett.*, to be published. doi: 10.1016/j.patrec.2018.10.002.
[60] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.
[61] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Weighted nuclear norm minimization with application to image denoising," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 2862–2869.
[62] Y. Chen and T. Pock, "Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1256–1272, Jun. 2017.
[63] H. C. Burger, C. J. Schuler, and S. Harmeling, "Image denoising: Can plain neural networks compete with BM3D?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 2392–2399.
[64] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Color image denoising via sparse 3D collaborative filtering with grouping constraint in luminance-chrominance space," in *Proc. IEEE Int. Conf. Image Process.*, San Antonio, TX, USA, Sep./Oct. 2007, pp. I-313–I-316.
[65] L. St-Laurent, D. Prévost, and X. Maldague, "Thermal imaging for enhanced foreground-background segmentation," in *Proc. Int. Conf. Quant. InfraRed Thermography*, Padova, Italy, Jun. 2006, pp. 27–30.

NA RAE BAEK received the B.S. degree in electronics and electrical engineering from Dongguk University, Seoul, South Korea, in 2017, where she is currently pursuing the M.S. and Ph.D. degrees in electronics and electrical engineering. She designed the gender recognition system based on CNN, analyzed results of experiments, and wrote the original paper. Her research interests include biometrics and pattern recognition.

SE WOON CHO received the B.S. degree in electronics and electrical engineering from Dongguk University, Seoul, South Korea, in 2017, where he is currently pursuing the combined M.S. and Ph.D. degrees in electronics and electrical engineering. He helped to perform the experiments and collect databases. His research interests include biometrics and pattern recognition.

JA HYUNG KOO received the B.S. degree in electronics and electrical engineering from Dongguk University, Seoul, South Korea, in 2016, where he is currently pursuing the combined M.S. and Ph.D. degrees in electronics and electrical engineering. He helped to collect databases. His research interests include biometrics and pattern recognition.



NOI QUANG TRUONG received the B.S. degree in computer engineering from the Hanoi University of Science and Technology, Hanoi, Vietnam, in 2016. He is currently pursuing the M.S. degree in electronics and electrical engineering from Dongguk University, Seoul, South Korea. He helped to port the algorithm on embedded systems. His research interests include biometrics and pattern recognition.

KANG RYOUNG PARK received the B.S. and M.S. degrees in electronic engineering and the Ph.D. degree in electrical and computer engineering from Yonsei University, Seoul, South Korea, in 1994, 1996, and 2000, respectively. He has been a Professor with the Division of Electronics and Electrical Engineering, Dongguk University, since March 2013. His research interests include image processing and biometrics. He has supervised this research and revised the original paper.